Introduction to Genome Analysis

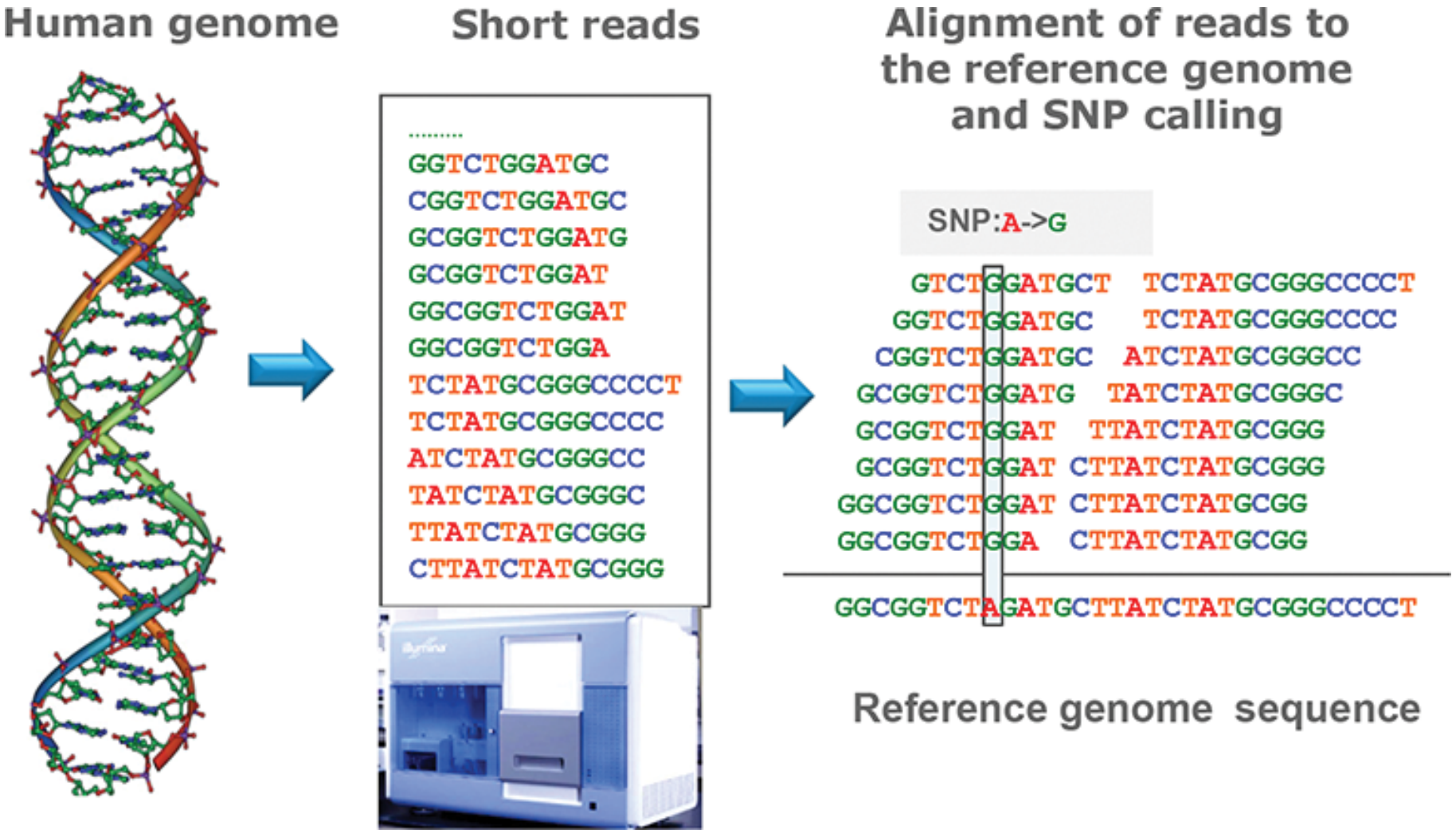
Christina Austin-Tse, PhD, FACMG
Clinical Molecular Geneticist, Center for Genomic Medicine, MGH
Instructor in Pathology, HMS
Analysis Team Lead, Rare Genome Project, Broad Institute of MIT and Harvard

CENTER FOR GENOMIC MEDICINE
MGH 1811 MASSACHUSETTS GENERAL HOSPITAL
Mass General Brigham
PERSONALIZED MEDICINE
LABORATORY FOR MOLECULAR MEDICINE
BROAD INSTITUTE

Roadmap

- A brief introduction
  - Next generation sequencing
  - Genome sequencing
- Genomic analysis
  - Data annotations
  - Data filtration strategies
  - Candidate follow-up/Reporting
- Limitations of genome sequencing

Next Generation Sequencing (NGS) generates short fragments of sequence that are aligned to a reference

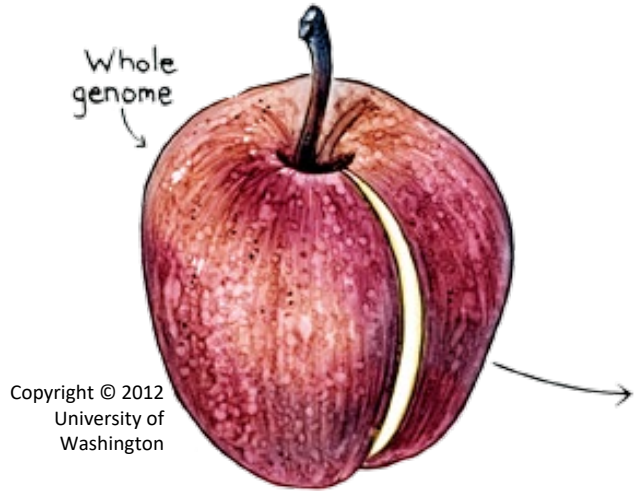


https://www.intechopen.com/books/cloud-computing-architecture-and-applications/cloud-computing-for-next-generation-sequencing-data-analysis

NGS can sequence the entire genome, or a small part of it

"Genomic Sequencing"

Genome Sequencing (GS)



- Virtually all of your DNA (~3 billion base pairs)
- Includes known disease genes, novel genes, and noncoding regions

Exome Sequencing (ES)

- ~21,000 protein coding genes (~30 million base pairs)
- Includes known disease genes and novel genes

Panel/ Targeted Sequencing

- Typically <100 genes
- Genes are selected for well-established role in disease

Genome sequencing can uncover many types of genetic variation

Well-established:

Single Nucleotide Variants (SNVs) CAGGTG → CACGTG

Small insertion/deletions (indels) CAGGTG → CAGTG or CAGGGTG

Trickier, but possible:

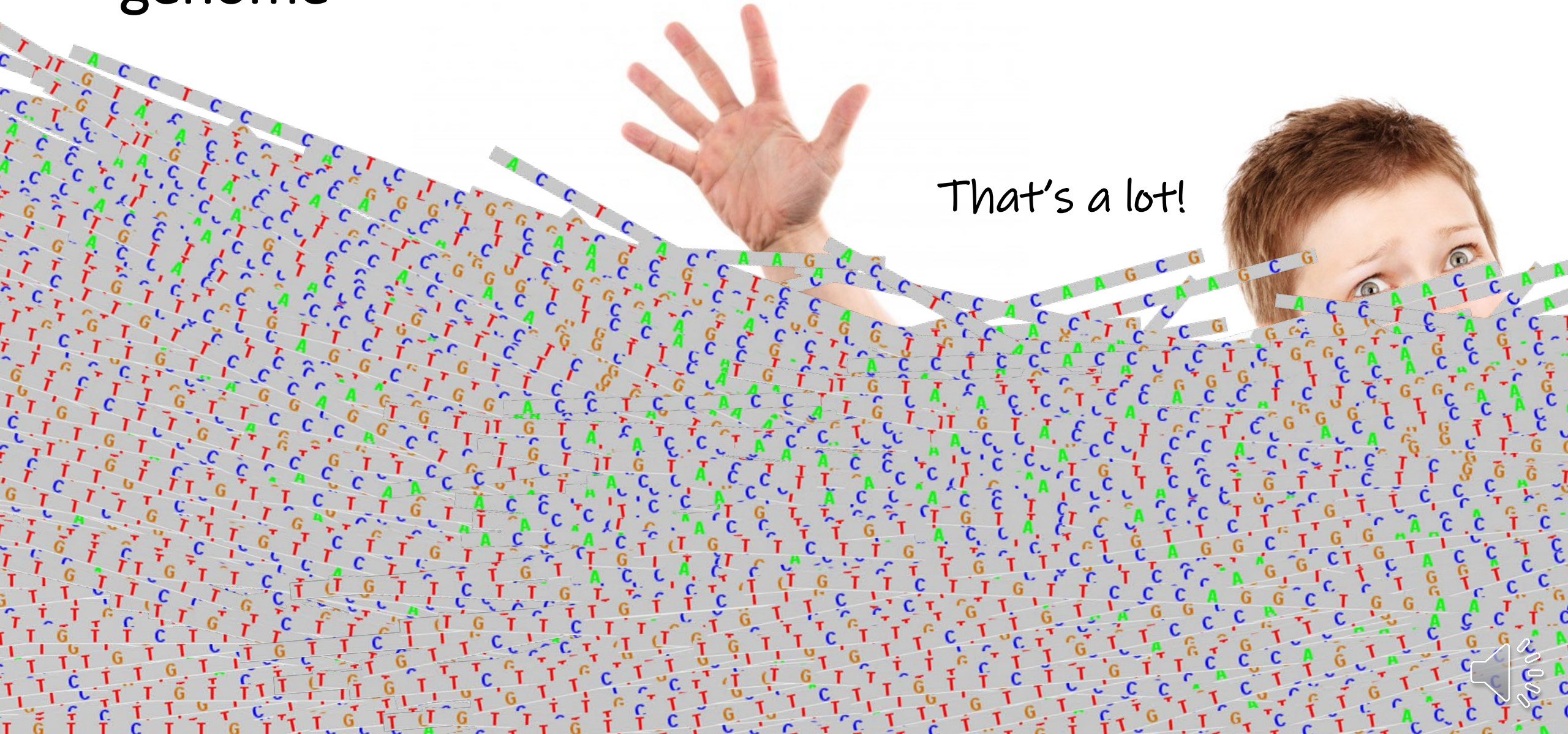
Short tandem repeats (STRs)/Repeat expansions (RE) CAGGTG → CAGCAGCAGCAG..... CAGGTG

Copy number variants (CNVs) – deletions, duplications

Balanced structural variants (SVs) – e.g. inversions

…and more

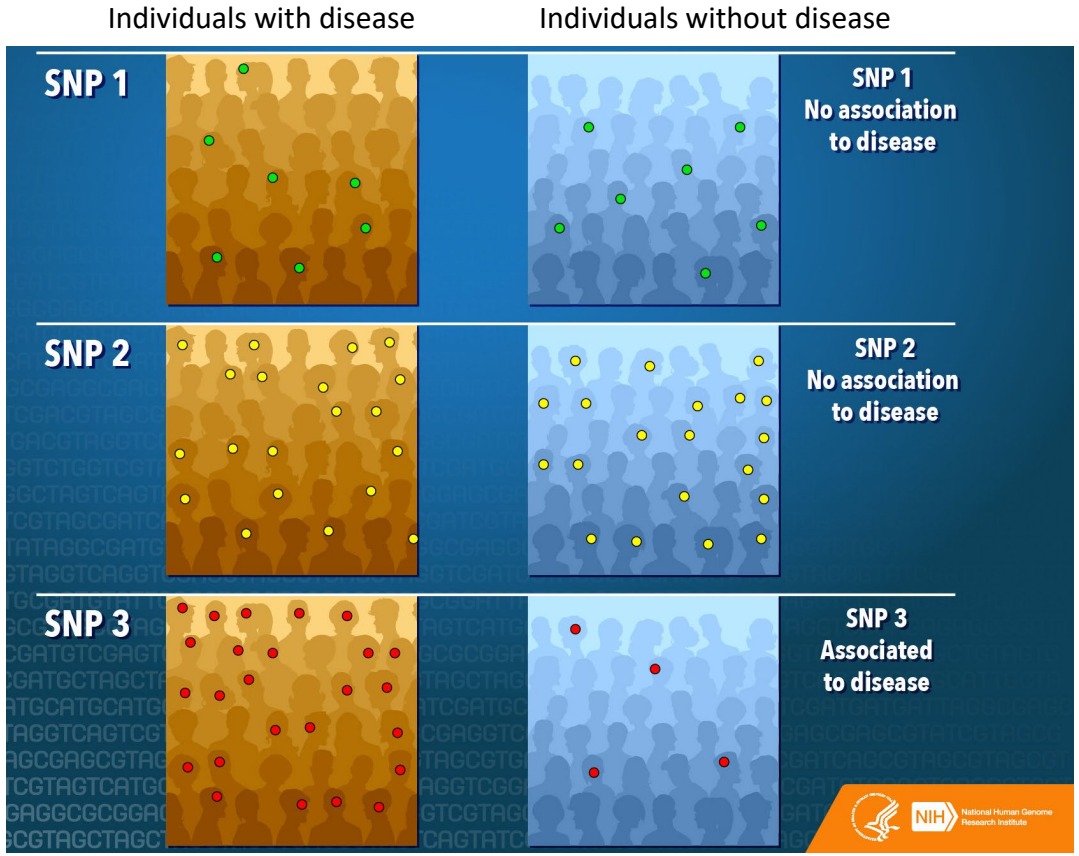
Millions of variants are identified in a typical genome



How do you find the disease-causing variants/genes?

Genomic sequencing data can be used in many research contexts

Common/complex disease: genome-wide association studies (GWAS)

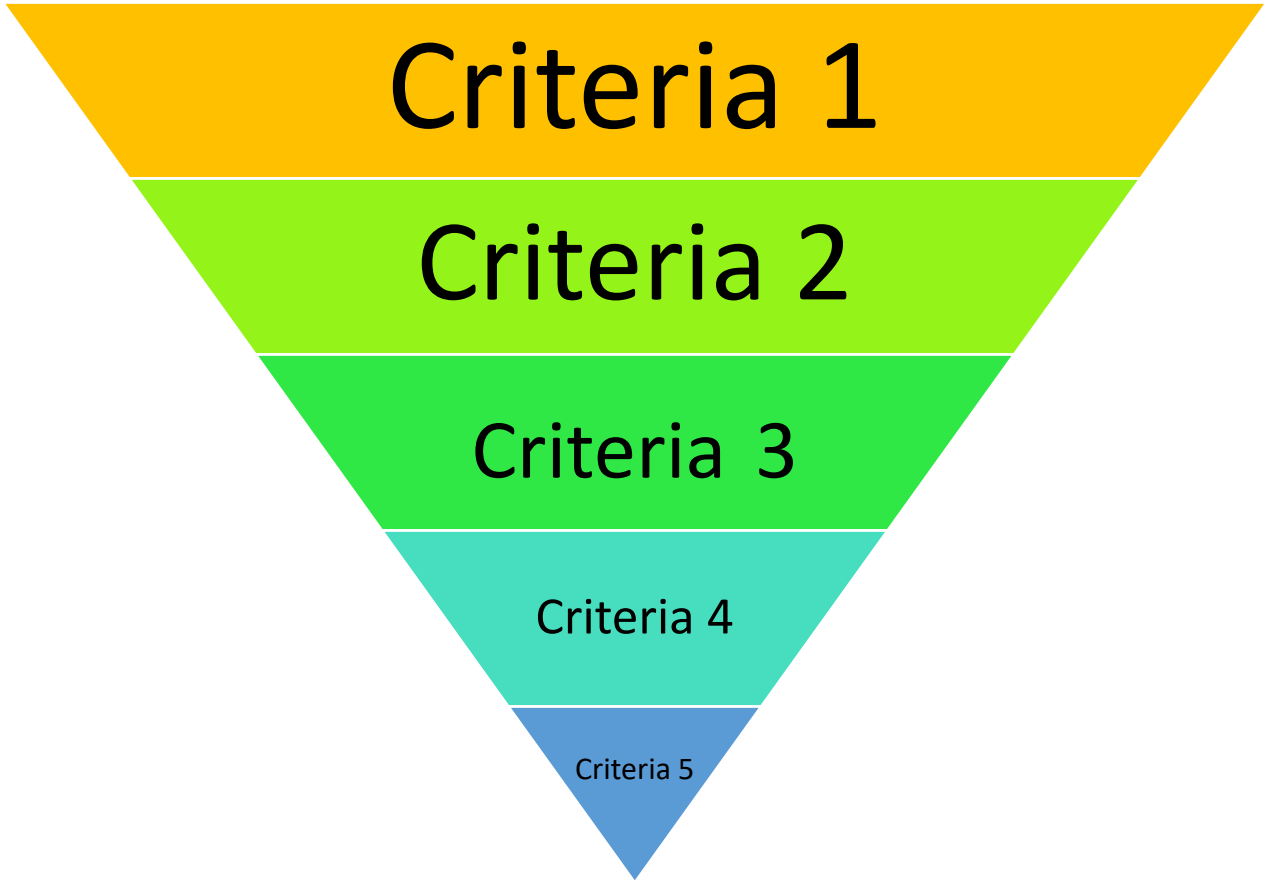


https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

Rare monogenic disease

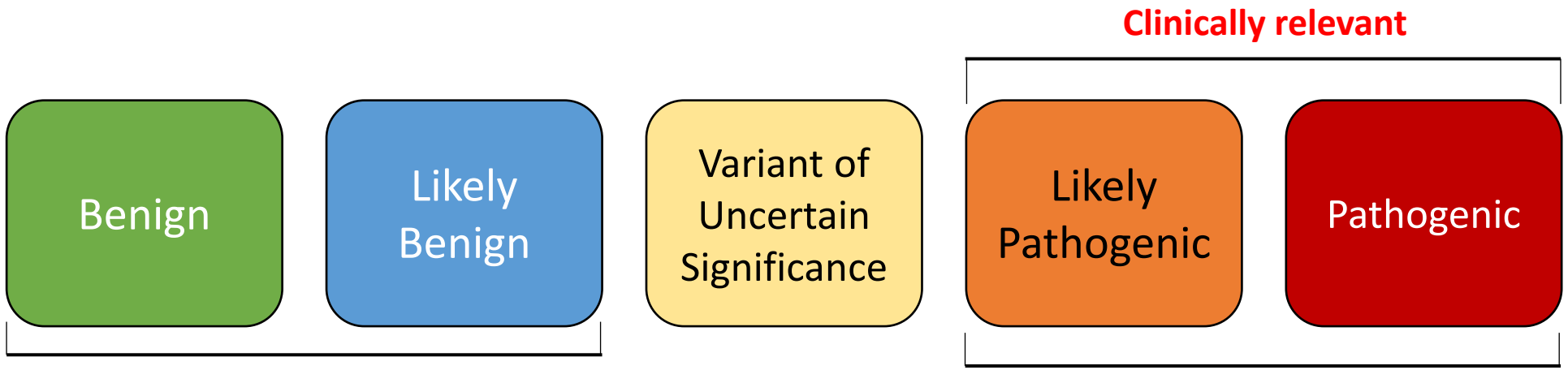
Filtration makes the work manageable

Millions of variants identified in a genome



Variants of interest

What variants might be of interest?



- Population frequency > disease prevalence
- No impact on amino acid sequence
- Changes amino acid at a poorly conserved position
- Inheritance not supportive of a disease causing role

- Rare
- Severe protein impact
- Reported in other individuals with consistent phenotypes
- Segregates with disease in families
- De novo occurrences
- Functional studies supportive of an impact

More detail on variant classification will be provided in Dr. Harrison's lecture

GS bioinformatic pipelines “annotate” variants with lots of information to enable filtration

- Quality Metrics
- Population Frequency
- Variant-disease associations
- Gene-level info
- Zygosity/ Inheritance
- Variant Impact

Case Example

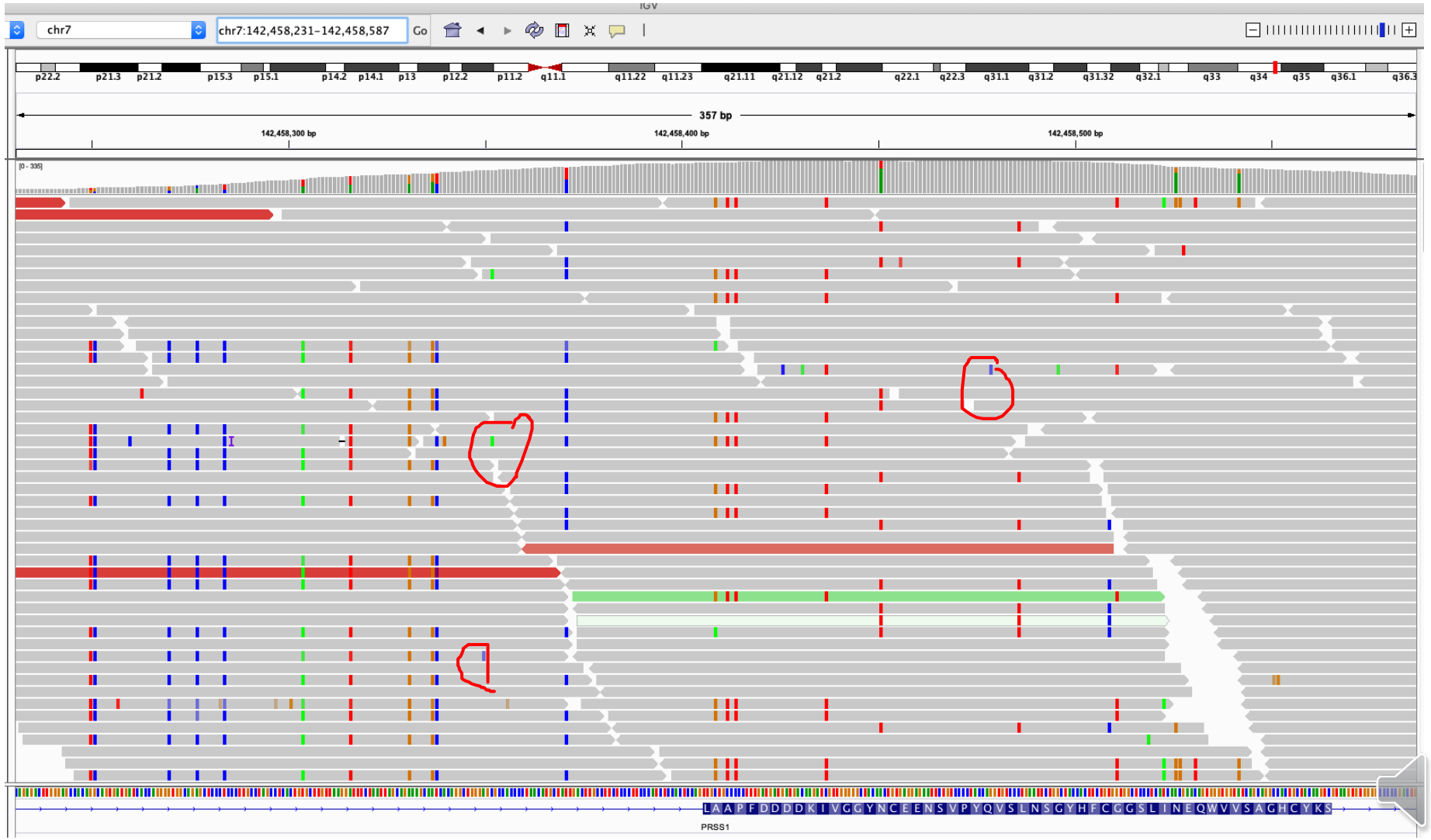
- 18 yo male with aplastic anemia, recurrent oral ulcers.
- Father, paternal grandmother, and 2 siblings also noted to have recurrent oral ulcers
- Whole exome sequencing ordered

67,000 Variants Identified

Variant Quality Metrics

Problem: Not all “variants” in the data set are real

Solution: Use quality metrics to remove spurious calls so we don’t waste time analyzing them



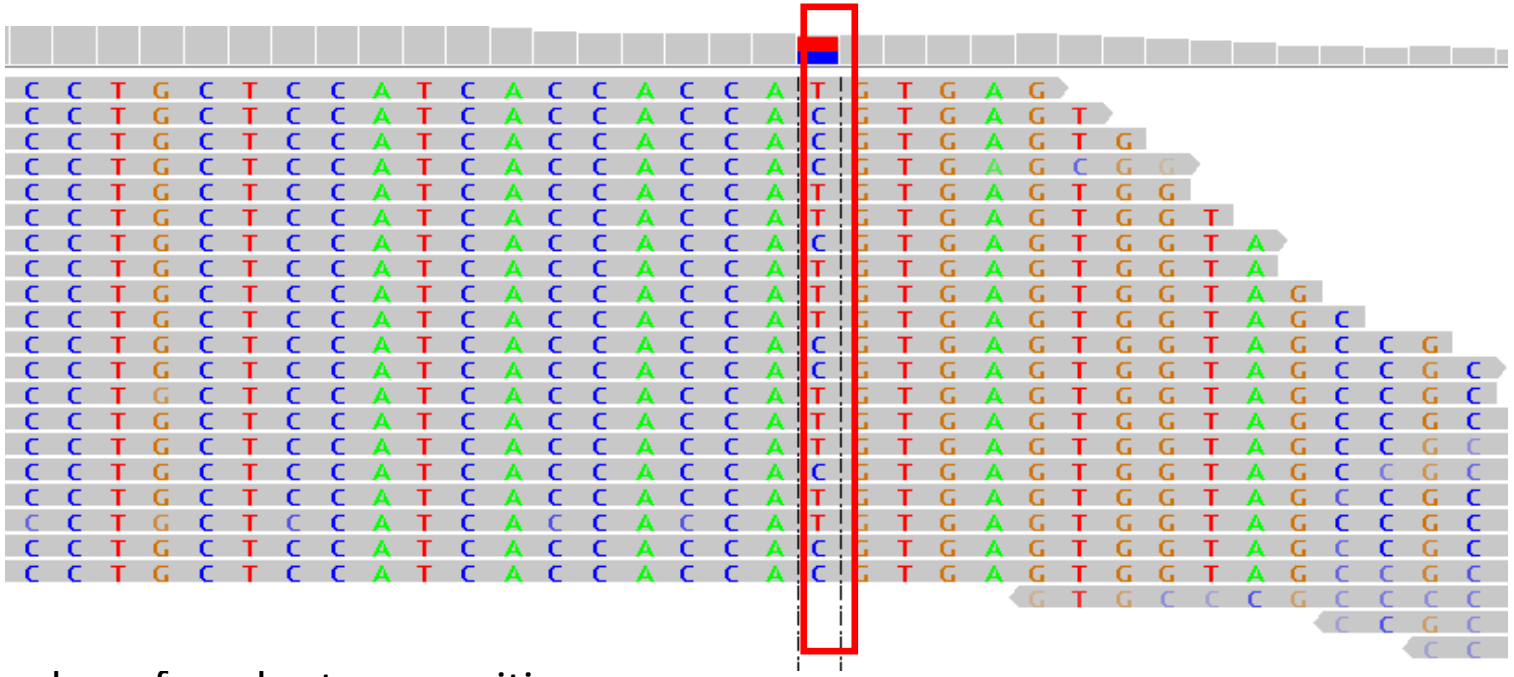
Common data quality metrics

Filter information based on defined numerical values:

- Allele balance – what fraction of the sequence fragments contain the variant? (Expect ~50% for a heterozygous change)
- Mapping quality – do the sequence fragments containing the variant line up unambiguously with the reference?
- Base quality – likelihood that the sequencer made the correct call at a position
- Depth of coverage – how much sequence information do we have at this position?

Good coverage is a requirement for accurate variant calling

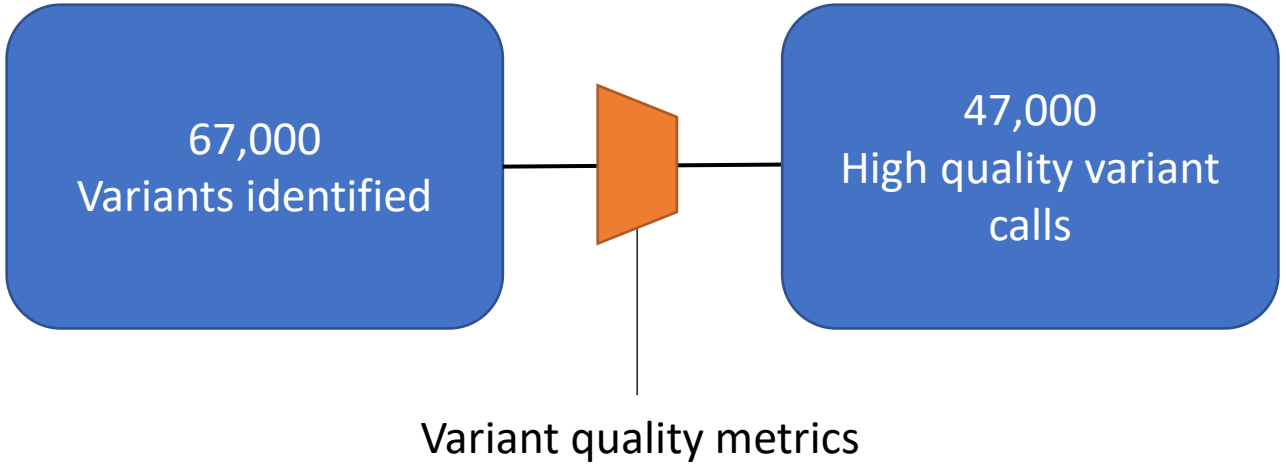
NGS sequencing data is <u>only</u> reliable when supported by a sufficient number of reads



- Coverage = The number of reads at any position
- Determines ability to reliably detect variants
- As amount of DNA being sequenced increases, coverage will be sacrificed (decreased)

Case Example

- 18 yo male with aplastic anemia, recurrent oral ulcers.
- Father, paternal grandmother, and 2 siblings also noted to have recurrent oral ulcers
- Whole exome sequencing ordered



Population Frequency

- **Problem**: We all carry a lot of variants that are too common to be disease-causing

```
                Reference sequence:  AGTGCCCATAGTGGA
60% of the general population: AGTGCCCATAGTGGA
40% of the general population: AGTGCCCTTAGTGGA
```

- **Solution**: Filter data based on frequencies in large population datasets (ie gnomAD)

gnomAD

genome aggregation database

http://gnomad.broadinstitute.org/

Exomes and genomes from >100,000 individuals (those with early onset Mendelian diseases excluded)

Population Frequencies

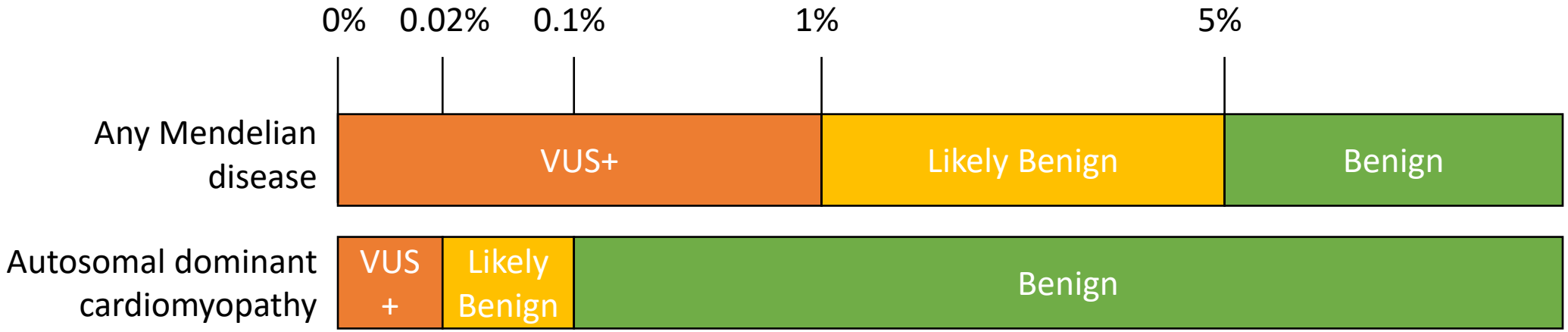
Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	8705	19640	1528	0.4432
European (non-Finnish)	23280	89200	1258	0.2610
European (Finnish)	5051	19396	310	0.2604
Other	1195	5318	66	0.2247
South Asian	3620	22254	133	0.1627
Ashkenazi Jewish	1056	6662	20	0.1585
Latino	4051	28046	132	0.1444
East Asian	343	15902	0	0.02157
Male	25822	111662	1815	0.2313
Female	21479	94756	1632	0.2267
Total	**47301**	**206418**	**3447**	**0.2292**

More detail on using gnomAD in Dr. O'Donnell-Luria's lecture

When is allele frequency too high to be pathogenic?

Depends on a number of factors:

- How common is the disease (prevalence)?
  - Consider founder effects
- What is the inheritance pattern?
- Does everyone with a pathogenic variant develop a phenotype (penetrance)?
- How many different genes/variants cause the disease

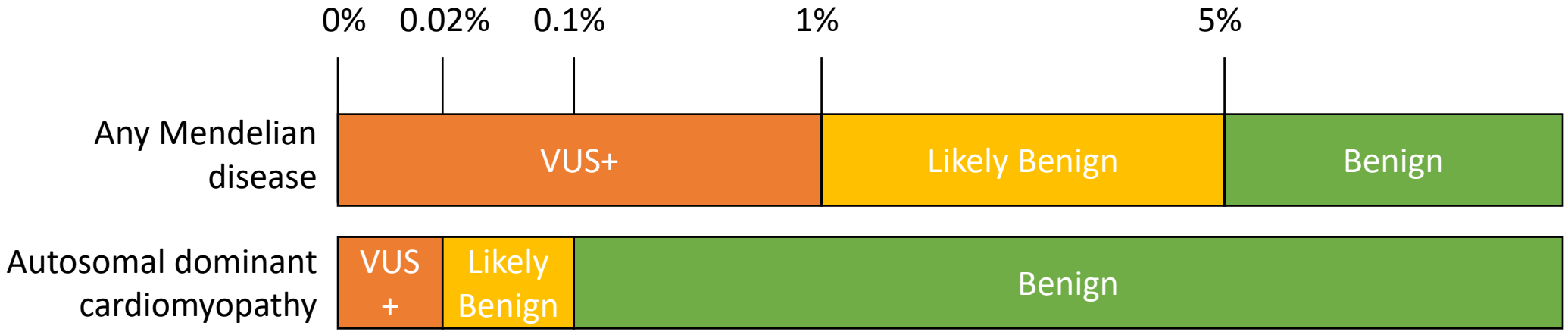


Visit http://www.cardiodb.org/allelefrequencyapp/ for a handy calculator

When is allele frequency too high to be pathogenic?

Depends on a number of factors:

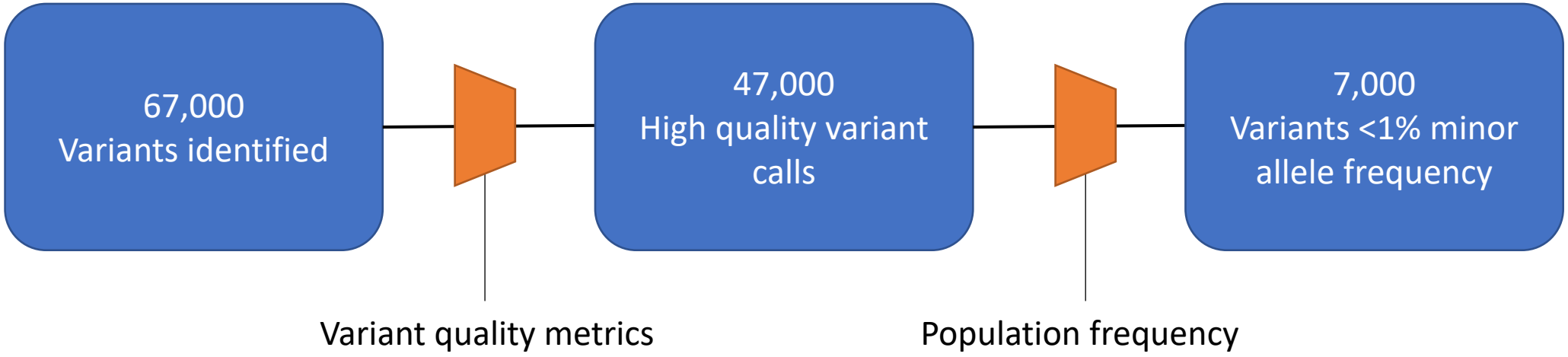
- How common is the disease (prevalence)?
  - Consider founder effects
- What is the inheritance pattern?
- Does everyone with a pathogenic variant develop a phenotype (penetrance)?
- How many different genes/variants cause the disease



Visit http://www.cardiodb.org/allelefrequencyapp/ for a handy calculator

Case Example

- 18 yo male with aplastic anemia, recurrent oral ulcers.
- Father, paternal grandmother, and 2 siblings also noted to have recurrent oral ulcers
- Whole exome sequencing ordered



Where do we go from here?

7,000
Variants <1% minor allele frequency

CHOOSE YOUR OWN ADVENTURE®

Genotype-centric

Show me the variants that are most likely to be classified as disease-causing

Phenotype-centric

Show me all variants in a list of genes that are relevant to my patient's phenotype

Inheritance-centric

For suspected recessive disease: Show me all homozygous or compound heterozygous variants

Gene-Disease Association

If there are known or candidate genes associated with your phenotype of interest, you may want to screen all variants in those genes

- **Problem:** How do I know which genes to look at?
- **Solution:** Assemble gene lists by integrating information from multiple databases

Variant databases:
- ClinVar
- HGMD
- Others

Gene-disease reference materials:
- GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/)
- Genetics Home reference (http://ghr.nlm.nih.gov/)
- OMIM (http://www.ncbi.nlm.nih.gov/omim)

Other places to look:
- PanelApp (http://panelapp.genomicsengland.co.uk/panels/)
- Genetic Testing Registry (http://www.ncbi.nlm.nih.gov/gtr/)
- PubMed (https://www.ncbi.nlm.nih.gov/pubmed)

NOTE: There is no single comprehensive source for gene lists!

More detail on assessing gene/disease associations in Erin Rooney Riggs' lecture

Where do we go from here?

7,000
Variants <1% minor allele frequency

CHOOSE YOUR OWN ADVENTURE®

Variant-centric

Show me the variants that are most likely to be classified as disease-causing

Phenotype-centric

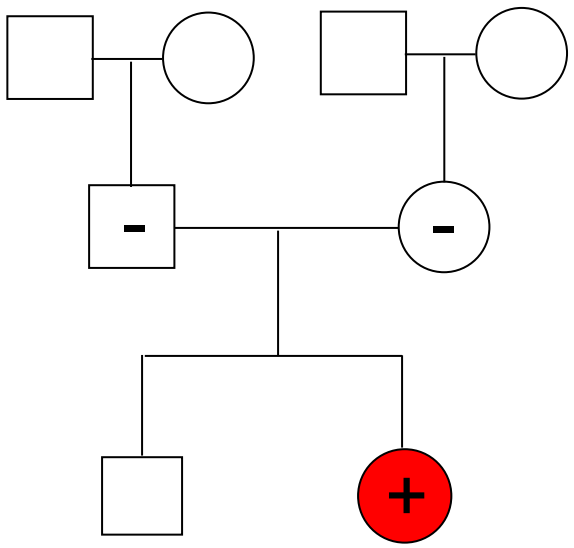
Show me all variants in a list of genes that are relevant to my patient's phenotype

Inheritance-centric

For suspected recessive disease: Show me all homozygous or compound heterozygous variants

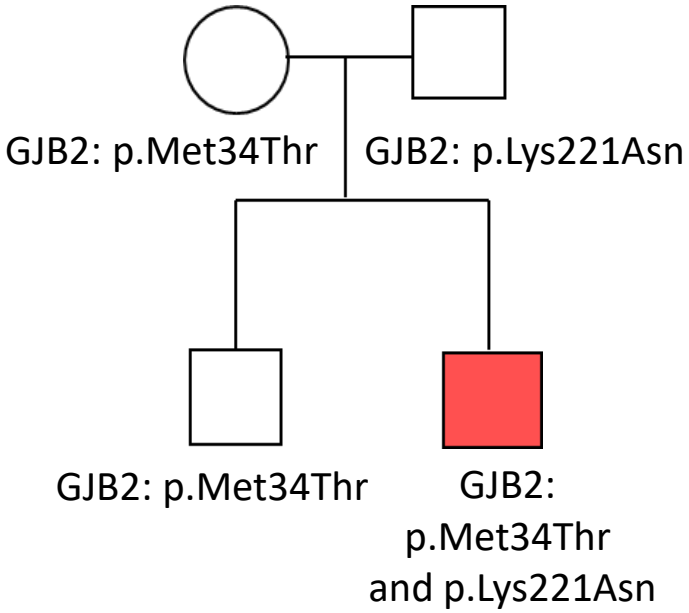
Inheritance Patterns

De Novo



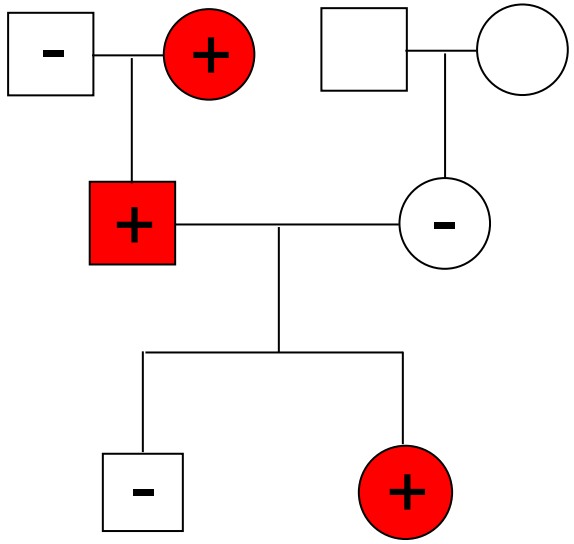
- Must have sequence data from parents
- Paternity/maternity should be confirmed

Recessive



- Screen for the presence of multiple rare variants in a single gene
- Can be done without WES/WGS data from parents

Dominant



- Filtrations are more powerful when you have sequence data from multiple affected family members
- Consider penetrance and expressivity

Where do we go from here?

7,000
Variants <1% minor allele frequency

CHOOSE YOUR OWN ADVENTURE®

Variant-centric

Show me the variants that are most likely to be classified as pathogenic

Phenotype-centric

Show me all variants in a list of genes that are relevant to my patient's phenotype

Inheritance-centric

For suspected recessive disease: Show me all homozygous or compound heterozygous variants

Variant Impact

Truncating
≠
Pathogenic

- Truncating Variants:
  - Nonsense – Any amino acid changed to a stop codon
  - Frameshift – Insertion or deletion of any number of bases that is not a multiple of 3, changes protein sequence and eventually introduces a stop codon
  - Splice variants – Changes to splice consensus nucleotides impact messenger RNA processing and subsequently protein sequence
- Other potentially interesting variants:
  - Start/stop loss
  - Missense (one amino acid changed to another)
  - Silent variants with splice impacts
  - Noncoding variants that impact RNA expression

Published data

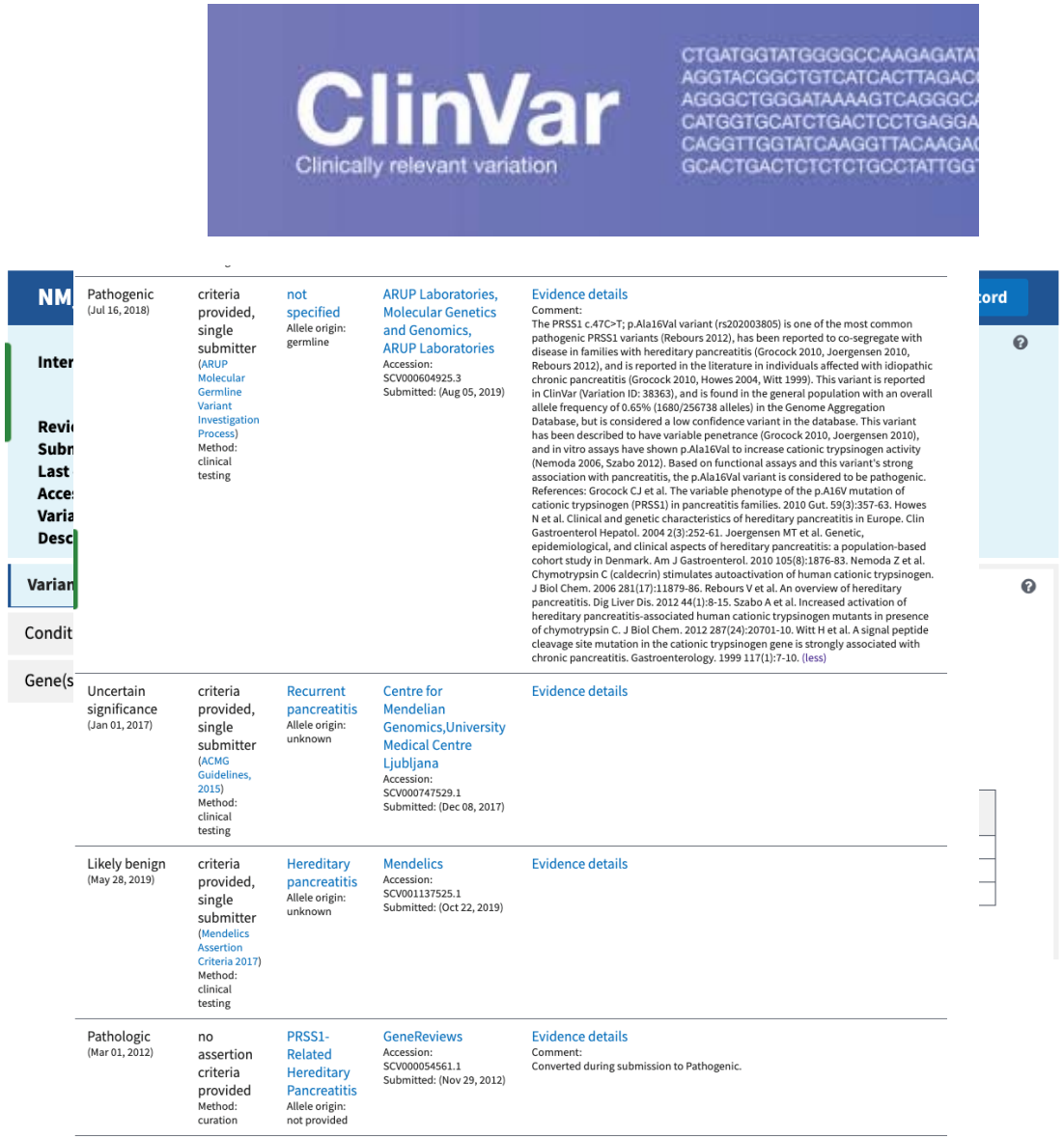
- **Problem:** Need an efficient way to find variants that have been reported in association with your phenotype of interest
- **Solution:** Filter variants based on presence/absence/classification from several databases***
  - Web-based variant databases
    - Public/freely accessible: ClinVar, LOVD, others
    - Subscription based: HGMD, Mastermind, others
  - Lab internal databases

***NOTE: There is currently no single comprehensive database!

ClinVar

Aggregation of user-submitted variant classifications

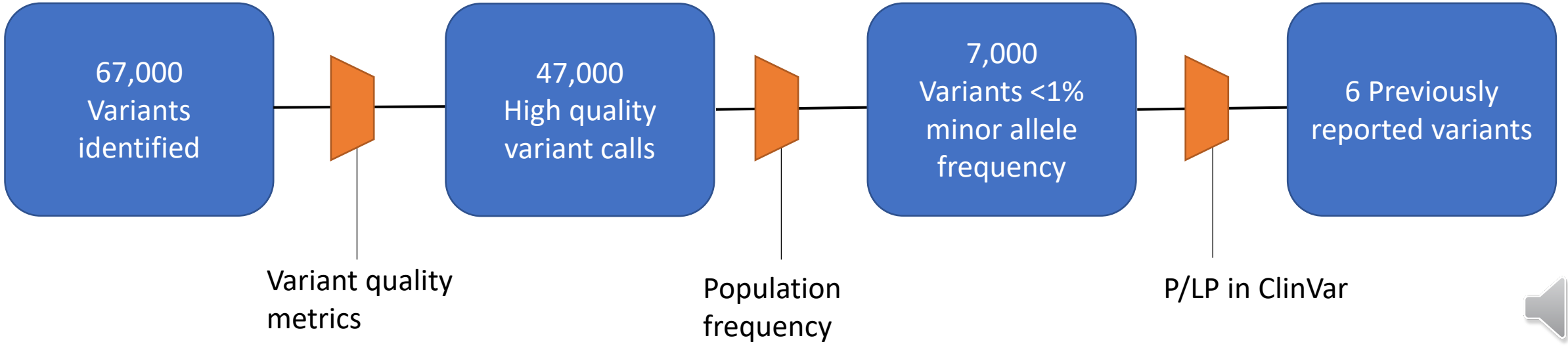
- https://www.ncbi.nlm.nih.gov/clinvar/
- Clinical, research, and expert panel classifications
  - Can have conflicts
  - Different levels of detail, accuracy
- Provides overall interpretation based upon individual classifications



More detail on using ClinVar in Dr. Landrum's lecture

Case Example

- 18 yo male with aplastic anemia, recurrent oral aphthous ulcers. Father, paternal grandmother, and 2 siblings also noted to have recurrent oral ulcers
- Whole exome sequencing ordered



Manual review of previously reported variants identifies a pathogenic change in STAT1

HGMD accession	Reported disease/phenotype	Variant class	Gene symbol	Codon change	Amino acid change	Codon number	Feedback
CM115709	Mucocutaneous candidiasis	DM	STAT1	GCG-GTG	Ala-Val	267	Feedback

The A267V substitution does not exhibit a shift in polarity and displays an increase in Kyte-Doolittle hydrophobicity from 1.8 to 4.2. Approximately 1.93% of missense mutations in HGMD are Ala-Val. The mutation occurs 484 amino acids from the end of the protein.

Literature citation	Citation type	Support	Comments/notes
1. van de Veerdonk (2011) *N Engl J Med* **365:** 54 PubMed: 21714643 STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis.	Primary literature report	✓	No comments
2. Sampaio (2013) *J Allergy Clin Immunol* **131:** 1624 PubMed: 23541320 Signal transducer and activator of transcription 1 (STAT1) gain-of-function mutations and disseminated coccidioidomycosis and histoplasmosis.	Functional characterisation	✓	Incr. STAT1 phos., delayed dephos., incr. DNA binding & t/activ., & incr. protein inhibitor interaction.
3. Mizoguchi (2014) *J Leukoc Biol* **95:** 667 PubMed: 24343863 Simple diagnosis of STAT1 gain-of-function alleles in patients with chronic mucocutaneous candidiasis.	Additional literature report		None
4. Nielsen (2015) *BMJ Case Rep* **2015:** bcr2015211372 PubMed: 26494717 A STAT1-gain-of-function mutation causing Th17 deficiency with chronic mucocutaneous candidiasis, psoriasiform hyperkeratosis and dermatophytosis.	Additional phenotype	✓	TH17 deficiency
5. Zheng (2015) *Eur J Immunol* **45:** 2834 PubMed: 26255980 Gain-of-function STAT1 mutations impair STAT3 activity in patients with Chronic Mucocutaneous Candidiasis (CMC).	Functional characterisation	✓	None
6. Depner (2016) *J Clin Immunol* **36:** 73 PubMed: 26604104 The Extended Clinical Phenotype of 26 Patients with Chronic Mucocutaneous Candidiasis due to Gain-of-Function Mutations in STAT1.	Additional literature report	✓	None
7. Breuer (2017) *Respir Med* **126:** 39 PubMed: 28427548 Autosomal dominant gain of function STAT1 mutation and severe bronchiectasis.	Additional phenotype	✓	Bronchiectasis, severe

Food for thought: Does this finding explain the patient's aplastic anemia? What could we be missing based on the filtration approach we took?

Submitted interpretations and evidence

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Pathogenic (Jun 05, 2018)	criteria provided, single submitter (Nykamp K et al. (Genet Med 2017)) Method: clinical testing	Immunodeficiency 31a Mycobacterial and viral infections, susceptibility to, autosomal recessive Immunodeficiency 31C Allele origin: germline	Invitae Accession: SCV000812325.1 Submitted: (Aug 29, 2018)	Evidence details Publications PubMed (8) Comment: This sequence change replaces alanine with valine at codon 267 of the STAT1 protein (p.Ala267Val). The alanine residue is highly conserved and there is a ... (more)
Pathogenic (May 04, 2018)	criteria provided, single submitter (LMM Criteria) Method: clinical testing	Chronic mucocutaneous candidiasis (Autosomal dominant inheritance) Allele origin: germline	Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine Accession: SCV000966986.1 Submitted: (Mar 21, 2019)	Evidence details Publications PubMed (9) Comment: The p.Ala267Val variant in STAT1 has been reported in >10 individuals with chron ic mucocutaneous candidiasis (CMC) and segregated with disease in 16 individuals from ... (more)
Pathogenic (Jun 01, 2013)	no assertion criteria provided Method: literature only	IMMUNODEFICIENCY 31C Allele origin: germline	OMIM Accession: SCV000044278.4 Submitted: (Nov 21, 2014)	Evidence details Publications PubMed (3)

Familial candidiasis

At least 35 *STAT1* gene mutations have been identified in people with familial candidiasis, an inherited tendency to develop infections caused by the *Candida* fungus (commonly called yeast infections). Most people with familial candidiasis have chronic infections of the skin, nails, and mucous membranes such as the lining of the mouth, collectively called chronic mucocutaneous candidiasis, beginning in early childhood. Some people with *STAT1* gene mutations have additional features such as increased susceptibility to other infections and an increased risk of autoimmune disorders, in which the immune system attacks the body's own tissues or organs. This combination of signs and symptoms caused by *STAT1* gene mutations is sometimes called immunodeficiency 31C.

I have a candidate, now what?

- In-depth analysis of the association between the STAT1 gene and MCC
  - ClinGen has guidelines for assessing gene-disease associations: https://www.clinicalgenome.org/docs/summary-of-updates-to-the-clingen-gene-clinical-validity-curation-sop-version-7/ (see Erin Rooney Riggs' lecture)
- Assessment of the Ala267Val variant according to ACMG/AMP variant classification criteria
  - PMID 25741868 (see Steven Harrison's lecture)
- Variant confirmation by an orthogonal technology
  - Requirement for this depends on setting, variant type

Sample filtration approaches: “Low hanging fruit”

Genotype-centric

Previously reported variants:
- ClinVar Pathogenic, Likely Pathogenic
- gnomAD popMAX <5%*

+

Predicted loss-of-function variants:
- Nonsense, Essential splice site, Frameshift, stop lost, initiator codon, start lost
- gnomAD popMAX <1%*
- Medically relevant genes

↓

Manual review

Phenotype-centric

Gene list-based filtration:
- Location: [Your list of selected genes]
- gnomAD popMAX <1%*

*Modify based on the phenotype in question

↓

Manual review

These approaches are used to identify variants in genes with established disease associations.

*Can modify based on the phenotype in question

What types of variants might be considered for return?

- Clinical exome and genome sequencing reports usually focus on Pathogenic and Likely Pathogenic variants related to the patient phenotype
  - Suspicious VUS may also be reported

Variant categories

- Rare, large effect variants that may explain the patient phenotype
- Variants that confer increased risk for the patient phenotype
- Medically actionable variants unrelated to the patient phenotype (incidental findings/secondary findings)
- Pharmacogenomic variants
- Carrier status

Return of variants that confer a diagnosis of a non-actionable diseases unrelated to the phenotype is generally discouraged in the clinical world

Sample filtration approaches: Discovery

De novo

De novo variants:

- Present in proband, absent in both parents
- gnomAD popMAX <0.1%
- Nonsense, missense, essential splice, extended splice, frameshift, inframe indels
- High quality calls

Recessive

- Compound heterozygous, homozygous, or hemizygous (X-linked)
- gnomAD popMAX <0.1%
- Nonsense, missense, essential splice, extended splice, frameshift, inframe indels
- High quality calls

Dominant

- Heterozygous calls
- Shared in affected individuals (if available)
- gnomAD popMAX <0.1%
- Nonsense, missense, essential splice, extended splice, frameshift, inframe indels
- High quality calls

These are restrictive searches. If no candidates are identified, criteria should be relaxed:

- Quality metrics
- Frequency thresholds
- Noncoding regions

What types of variants might be considered for return?

- In research/discovery settings, additional data is often needed to evaluate whether or not a candidate is truly causative for the phenotype. These may include:
  - Transcript analysis
  - Functional studies – in vitro or in vivo
  - Identifying additional patients with variants in the same gene
    - Matchmaker Exchange (www.matchmakerexchange.org)
  - Additional clinical followup

Limitations of GS

- Not all variants are detectable by WGS
  - Highly homologous regions
  - Challenging variant types (repeat expansions, structural variants) may be missed depending on the pipeline in use
- Variants in noncoding regions are difficult to interpret
  - Typically requires supplementary assays, ie RNASeq
- Success of filtration methods depends on current understanding of genes, variants, and pathways
  - Consider periodic reinterpretation of unsolved cases

Take-home messages

- Genomic sequencing produces a lot of data
- Bioinformatic data processing and specialized filtration programs are essential to analysis
- There are many approaches to genomic data filtration
  - Different methods will return different subsets of variants
  - Different contexts necessitate different filtration approaches
- Analysis does not end at filtration
  - In depth assessment of genes and variants are necessary to determine if your candidates may be disease causing
  - Follow-up studies may be necessary

Good luck in your analyses!